Improved small molecule identification through learning combinations of kernel regression models

Celine Brouard, Florence d'Alché-Buc, Juho Rousu

Improved small molecule identification through learning combinations of kernel regression models

Céline BROUARD[1], Florence D'ALCHÉ-BUC[2] and Juho ROUSU[3]
[1] MIAT, UR 875, INRA, 31326, Castanet Tolosan, France
celine.brouard@inra.fr
[2] LTCI, Télécom Paris, Institut Polytechnique de Paris, 75634, Paris, France
florence.dalche@telecom-paris.fr
[3]HIIT, Department of Computer Science, Aalto University, 00076 Espoo, Finland
juho.rousu@aalto.fi

An important problem in the field of metabolomics is the identification of the metabolites present in a biological sample. Information on metabolites can be obtained using tandem mass spectrometry. This technology allows to obtain a tandem mass spectrum, also called MS/MS spectrum, by fragmenting a metabolite. In recent years, the massively increased amounts of publicly available reference MS/MS spectra in databases have caused a revolution in small molecule identification. In particular, the use of modern machine learning approaches has become feasible and lead to the generation of a host of machine learning approaches and identification tools such as FingerID [1], CFM-ID [2], CSI:FingerID [3], CSI:IOKR [4], SIMPLE [5]. The majority of the machine learning methods rely on the same conceptual scheme introduced with FingerID: predicting molecular fingerprints from tandem mass spectrometry data and finding the most similar fingerprint from a molecular structure database. This approach has been very successful, for example, CSI:FingerID and CSI:IOKR have been top performers in the most recent CASMI contests. The alternative conceptual approach for small molecule identification, sometimes called *in silico fragmentation*, calls for predicting MS/MS spectra for a set of candidate molecular structures and choosing the most similar predicted MS/MS spectrum to the observed MS/MS spectrum. This approach is used, e.g., in the non-machine learning based MetFrag [6] as well as CFM-ID [2] which is the most notable machine learning tool relying on the in silico fragmentation approach.

CSI:IOKR predicts the molecular structures through a single structured output prediction algorithm, called Input-Output Kernel Regression (IOKR) [7]. The principle of this method is to encode the similarities in the input (spectra) space and the similarities in the output (molecule) space using two kernel functions. This method approximates the spectra-molecule mapping in two steps. The first step corresponds to a regression problem from the input space to the feature space associated with the output kernel. The second phase is a preimage problem, consisting in mapping back the predicted output features to the molecule space. Due to this approach, CSI:IOKR is extremely fast to train and is on par with CSI:FingerID in accuracy. Both CSI:FingerID and CSI:IOKR make use of multiple data sources, fused using the multiple kernel learning (MKL) algorithm ALIGNF that sets importance weights to the input kernels prior learning the fingerprint prediction models.

In this work, we bring forward two methodological contributions. Firstly, we extend the IOKR approach by formulating essentially an in silico fragmentation problem which we call IOKRreverse. From a set of candidate molecular structures, we implicitly predict a representation of an MS/MS spectrum for each candidate, and solve a pre-image problem to output the molecular structure whose predicted MS/MS is the closest to the observed one. All this computation is done through kernel matrices of the inputs (MS/MS spectra) and outputs (molecular structures). Secondly, we introduce an approach called IOKRfusion to combine multiple IOKR and IOKRreverse models, which arise from the use of different input and output kernels on the training data. The models are combined by minimizing the structured Hinge loss, which is frequently used in structured output learning. This way of aggregating multiple data sources is sometimes called *late fusion*, since the model learning happens before the aggregation, as compared to multiple kernel learning using, which happens before model learning, making it an *early fusion* approach.

In the numerical experiments, we used two subsets of MS/MS spectra from GNPS and MassBank to evaluate

the performance of our method. The first subset contains 6974 MS/MS spectra measured with a positive ionization mode while the second subset contains 3578 MS/MS spectra measured with a negative ionization mode. The IOKRreverse and IOKRfusion models have been compared to CSI:IOKR and CSI:FingerID. Our experiments show a consistent improvement of the identification results on both datasets when using the IOKRfusion approach, showing the potential of the approach. In particular, we note that the late fusion approach, used by IOKRfusion, improves over the early fusion approach, which was used in CSI:IOKR. In addition, the IOKRfusion approach turned out to outperform CSI:FingerID. IOKRfusion is also fast to train and test. In conlusion, IOKRfusion can be seen to maintain the computational efficiency of the IOKR framework, while improving the small molecule identification accuracy.

References

[1] M. Heinonen, H. Shen, N. Zamboni and J. Rousu. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*. Vol. 28, pp : 2333-2341, 2012.

[2] F. Allen, A. Pon, M. Wilson, R. Greiner and D. Wishart. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res*. Vol. 43, pp. : W94-W99, 2014.

[3] K. Dührkop, H. Shen, M. Meusel, J. Rousu and S. Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *PNAS.* Vol. 112, pp. :12580-12585, 2015.

[4] C. Brouard, H. Shen, K. Dührkop, F. d'Alché-Buc, S. Böcker and J. Rousu. Fast metabolite identification with Input Output Kernel Regression. *Bioinformatics*. Vol. 32, pp. : i28-i36, 2016.

[5] D.H. Nguyen, C.H. Nguyen and H. Mamitsuka. SIMPLE: Sparse Interaction Model over Peaks or moLEcules for fast, interpretable metabolite identification from tandem mass spectra. *Bioinformatics*. Vol. 34, pp : i323-i332, 2018.

[6] C. Ruttkies, E.L. Schymanski, S. Wolf, J. Hollender and S. Neumann. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminform*. Vol. 8 (3), 2016.

[7] C. Brouard, M. Szafranski and F. d'Alché-Buc. Input Output Kernel Regression: supervised and semi-supervised structured output prediction with operator-valued kernels. *J. Mach. Learn. Res.* Vol. 17, pp : 1-48, 2016.